Trabajo de Fin de Grado

Herramientas estadísticas básicas aplicadas a la investigación sociolingüística

Autor: Ignacio Valdés Zamudio

Tutor: Luis Escoriza Morera

Grado en Lingüística y Lenguas Aplicadas

Curso académico 2015-2016

Convocatoria de junio

Facultad de Filosofía y Letras

Índice

Resumen

Resumen

Presentamos un trabajo interdisciplinar en el que estadística, computación, sociología y lingüística se combinan para proporcionar al investigador en sociolingüística herramientas de trabajo que le permitirán enriquecer sus estudios, así como sistematizar su proceder y, finalmente, obtener conclusiones científicas basadas en sólidas teorías matemáticas, con la ayuda del tratamiento informático, que le permitirá procesar enormes cantidades de información. El enfoque de nuestro trabajo es eminentemente práctico, procurando una asimilación gradual de los conceptos y herramientas de trabajo que consideramos básicas en la formación de un sociolingüista.

Palabras clave: 'estadística', 'computación', 'sociología', 'lingüística', 'sociolingüística'.

Abstract

We present an interdisciplinary work in which statistics, computing, sociology and linguistics are combined in order to provide the researcher in sociolinguistics working tools which will allow him to enrich his studies and systematize their actions and, finally, obtain scientific conclusions based on solid mathematical theories, with the help of computing processing, allowing him to process huge amounts of information. The focus of our work is eminently practical, seeking a gradual assimilation of the concepts and working tools that we consider basic in the training of a sociolinguist.

Keywords: 'statistics', 'computing', 'sociology', 'linguistics', 'sociolinguistics'.

Introducción

Antecedentes y estado del arte

Desde la eclosión de numerosas disciplinas lingüísticas en los años sesenta del pasado siglo XX, aunque con excepciones como, por ejemplo, la Teoría de la Información de Shannon y Weaver (1949), que supuso toda una revolución por poner en conexión cosas tan aparentemente diferentes como la lingüística y las matemáticas, explicando las relaciones, proporciones y leyes matemáticas que rigen la transmisión, procesamiento y medición de información, y cuyas aplicaciones han sido tan vastas en campos tan heterogéneos como los medios de comunicación o la criptografía, la lingüística no ha cesado de aproximarse a otras disciplinas o ramas del saber para nutrirse mutuamente y crear nuevos campos y vías de estudios interdisciplinares que aúnan las preocupaciones de las dos -o más- disciplinas de partida. Ejemplos de ello, serían: la sociolingüística (unión de sociología y lingüística), la psicolingüística (unión de psicología y lingüística), la lingüística forense (investigación forense basada en muestras lingüísticas), la lingüística computacional (desarrollo de herramientas lingüísticas y potenciación de las capacidades investigadoras mediante herramientas informáticas), etc.

En el contexto concreto de la investigación sociolingüística, existen -si bien no numerosos- manuales específicos o semiespecíficos que tratan el asunto. En la mayoría de ellos encontramos información semejante y con la misma opinión al respecto; veamos, por ejemplo, lo que Fasold (1984, pp.141-142) comenta en relación a esta cuestión:

> *La estadística es (...) un instrumento que se puede utilizar para sacar algo en claro de una gran cantidad de números. (...) Una vez que se han reunido los datos numéricos, no siempre resulta obvio ni siquiera qué significan. En estadística existen métodos para que el investigador sepa si los datos apoyan la idea que está tratando de probar. Este uso de la estadística se conoce con el nombre de estadística inferencial. (...) Hay que comprender los rudimentos de la estadística para hacer un estudio profundo de la sociolingüística.*

Y no solo se hace referencia a la importancia de la estadística, sino también, como será común en el desarrollo de nuestro trabajo, a la creciente relevancia que la computación está cobrando en el ámbito de las investigaciones lingüísticas, como menciona Mairal Usón,

Peña Cervel, Cortés Rodríguez, y Ruiz de Mendoza Ibáñez (2010, p.56):

> *Aunque tener en cuenta los juicios del hablante ha ayudado mucho al progreso de los modelos sobre el lenguaje, en tiempos más recientes la investigación lingüística se ha beneficiado del análisis de grandes corpus informatizados. Un corpus informatizado permite efectuar búsquedas automatizadas de palabras, sintagmas, categorías lingüísticas básicas (nombre, adjetivo, verbo, adverbio, preposición), colocaciones (palabras que aparecen en asociación frecuente unas con otras, como pan y queso) y los contextos en que aparecen los ítems buscados.*

En sus aplicaciones, que son tantas como podamos imaginar, estos autores dejan bien claro con ejemplos específicos cómo sacar partido a la estadística aplicándola a estudios sociolingüísticos concretos, como vemos en López Morales (1994, p.137):

> *En una investigación dada, el hecho elemental o unidad estadística podría ser el segmento fonológico subyacente -/s/; sus diferentes realizaciones de superficie serían las variantes ([s,h,0]), mientras que los factores lingüísticos vendrían a ser, por ejemplo, 1) la posición en la palabra (sílaba final o interior), 2) el segmento fonológico siguiente (vocal, consonante, no segmento) y 3) la categoría gramatical del segmento. Los factores extralingüísticos coinciden generalmente con las variables de la investigación: procedencia, edad, sexo, etc.*

Normalmente, encontramos en estos autores y manuales información general sobre cuestiones estadísticas y/o metodológicas, nunca desarrolladas en el caso de la matemática/estadística, como podemos ver en Piergiorgio (2007, p.146):

> *Una encuesta por muestreo es un modo de obtener información: a) preguntando, b) a los individuos que son objeto de la investigación, c) que forman parte de una muestra representativa, d) mediante un procedimiento estandarizado de cuestionario, e) con el fin de estudiar las relaciones existentes entre las variables.*

O en López Morales (1994, p.41):

> *A menos de que se trate de un universo muy pequeño -una familia, una aldea, los alumnos de un aula, de un colegio, los obreros de una fábrica, etc.- es imposible trabajar con técnica de censo, es decir, obtener información de todos y cada uno de los individuos que integran la población.*

Y, por último, hasta podemos encontrarnos pequeñas discusiones sobre ventajas y desventajas de una técnica específica, sobre inconvenientes o problemas que pueden presentar determinadas herramientas estadísticas, etc., como en Piergiorgio (2007, p.112):

> *El investigador tiene dos modos de asegurarse de que la relación entre X e Y no se debe a la acción de Z sobre las dos variables: a) el control, es decir, la transformación de las variables extrañas en constantes, y b) la depuración, es decir, la determinación por procedimientos matemáticos y la consecuente eliminación de los efectos de las variables extrañas.*

O, de nuevo, en López Morales (1994, p.112):

> *Las preguntas cerradas tienen la virtud de preguntar directamente lo que se quiere saber. La experiencia ha demostrado que este tipo de pregunta es mejor y más fácilmente respondida por el sujeto que las abiertas. Pero, además, poseen el mérito de facilitar la revisión del cuestionario, sobre todo cuando se trabaja cuantitativamente, ya que la tabulación de los resultados se convierte en un proceso relativamente simple y cómodo. (. . .) Las preguntas abiertas son más fáciles de preparar, pero requieren un proceso de revisión mucho más complicado. Algunos autores las prefieren para cierto tipo de investigaciones porque suponen que no influyen en la respuesta, sobre todo cuando se conciben en términos muy generales. Pero, además de los inconvenientes de tabulación, se enfrentan a otros dos contratiempos; primero, que están estrechamente relacionadas con la capacidad expresiva del sujeto, es decir, que si nuestro informante procede de niveles culturales más aventajados logrará transmitirnos su pensamiento de manera adecuada, pero no será así siempre si el sujeto carece de las destrezas necesarias para ordenar y expresar con claridad sus ideas. Por otra parte, el tener que elaborar su respuesta sin ningún (o con poco) apoyo por parte de la pregunta puede hacerlo caer bajo la influencia del entrevistador; las respuestas pueden variar con el comportamiento del que pregunta.*

En resumen, a pesar de que hay algunos autores que han tratado la cuestión de la metodología en el campo de la sociolingüística, es complicado encontrar un manual donde se explicite la comprensión y uso de las herramientas estadísticas; tal vez, el que más se aproxime sea Hernández Campoy y Almeida (2005), si bien no explica detalladamente cómo abordar las cuestiones de procedimiento estadístico ni plantea un tratamiento computacional desarrollado

más allá de la mención de programas que realizan estas tareas de forma automática, pero en los cuales el investigador no tiene ningún control.

La situación actual presenta, pues, una menor interdisciplinariedad de la metodología en el área de la sociolingüística en relación a otras disciplinas de la lingüística, como pueden ser la lingüística computacional, la ingeniería del lenguaje, etc. Asimismo, echamos en falta una mayor formación en estadística en los estudios actuales de lingüística, que permita tanto al alumnado como a los investigadores poder dar un caracter más científico y riguroso, así como sistemático, a sus trabajos, estudios o investigaciones. Dicho de otro modo, queremos dar cuenta de la falta de conexión entre disciplinas no solo compatibles, sino necesarias, que encontrarían su confluencia, sin lugar a dudas, en el campo de la sociolingüística, a saber: estadística, computación, sociología y lingüística.

Objetivos

Presentamos una visión interdisciplinar dentro del marco de una disciplina que ya es interdisciplinar por sí misma. Nuestro trabajo versa sobre cómo usar -y no solo comprender- herramientas estadísticas, mediante el uso de la computación, en la resolución de problemas sociolingüísticos que nos permitan extraer conclusiones más jugosas, rigurosas y sistemáticas; es decir, pretendemos aunar la ciencia estadística, con datos cuantitativos (en tanto que vertiente metodológica), con la computación (herramienta concreta de trabajo) para, a su vez, aplicarlo a una disciplina (sociolingüística) que es mezcla de otras dos: sociología y lingüística.

Nuestro objetivo es doble: por un lado, llamar la atención sobre la necesidad de incluir estas disciplinas en el currículo formativo de un investigador en sociolingüística, así como mostrar de manera específica y sistemática la interconexión que se establece entre todas las áreas mencionadas; por otro lado, ofrecer una propuesta de recorrido formativo en el que todas las disciplinas entran en acción de manera conjunta para que la interdisciplinariedad quede puesta en valor y justificada sobradamente. Aunando ambos objetivos, podemos concluir en que el propósito último de nuestro trabajo consiste en facilitar la penetración, a través de la demostración de su conveniencia y oportunidad, de la ciencia estadística, en íntima unión con la computación, en los estudios lingüísticos actuales en general, y sociolingüísticos en particular.

Metodología y organización

Intentando ser fieles al objetivo de nuestro trabajo, hemos buscado la sistematicidad, organización y claridad en la exposición de los contenidos. Todos los capítulos y apartados, a excepción del primero referido a la Población y Muestra, por tener un carácter meramente informativo y no de tratamiento cuantitativo, presentan la misma estructura: definición formal de la herramienta estadística tratada, breve explicación o aclaración de la misma, ejemplificación, apoyada en bibliografía, de posibles aplicaciones de la herramienta, creación y resolución, a través de herramientas computacionales, de un ejercicio propio basado en la bibliografía mencionada.

Para el seguimiento apropiado del hilo conductor que proponemos, suponemos conocimientos básicos en sociolingüística y programación, pues forman parte del itinerario formativo de la práctica totalidad de grados universitarios en Lingüística actuales. En relación a las cuestiones estadísticas, hemos procurado la autosuficiencia; es decir, nuestro propósito ha sido evitar la necesidad de conocimientos estadísticos previos, por lo que comenzaremos por las cuestiones más básicas y elementales de esta ciencia matemática, para ir progresivamente aumentando la complejidad de los conceptos y herramientas utilizadas, pretendiendo con ello facilitar un aprendizaje gradual, en el caso de que sea necesario, de las cuestiones tratadas.

Evidentemente, nuestro trabajo no abarca la totalidad de herramientas disponibles para la investigación en sociolingüística, ni la información que proporcionamos de las tratadas son exhaustivas, pues nuestro objetivo no es convertir a los sociolingüistas en estadísticos, sino que consigan autonomía en el desarrollo de sus labores, sin depender del trabajo o herramientas proporcionadas por otros.

Finalmente, es preciso destacar la falta de bibliografía específica al respecto y ello es lo que explica la carencia de un aparato bibliográfico mayor, así como la falta de necesidad de recurrir a multitud de manuales para cuestiones como, por ejemplo, dar las diferentes definiciones de las herramientas estadísticas utilizadas. Nuestro trabajo no pretende ser, por tanto, una investigación teórica basada en la toma del testigo de numerosos autores previos, sino una propuesta completamente práctica de ampliación de horizontes en cuanto a la interdisciplinariedad que, a nuestro entender, debe existir en el campo de la sociolingüística.

Población y muestra

Según López Morán y Hernández Alonso (2012, p.18),

> *toda investigación estadística, sea cual sea su tipo y finalidad, debe tener como referencia última un conjunto de personas o cosas a las que va dirigida la investigación. Este conjunto, marco obligado de referencia para cualquier estudio estadístico, se conoce como* ***población****.*

Por otro lado, también conforme a López Morán y Hernández Alonso (2012, p.22),

> *se habla de* ***muestra*** *cuando [...] solo se investiga una parte de la población. [...] es preciso tomar los elementos que constituyen la muestra de forma que sean plenamente representativos del conjunto de elementos que conforman la población.*

Como el lector puede estar pensando, la selección de dicha muestra condicionará los resultados del estudio estadístico y, por tanto, su extrapolación a la población en general. Es por ello que un buen investigador debe tener la formación suficiente para no cometer errores metodológicos a la hora de llevar a cabo el muestreo. Como norma básica, se debe recurrir a procedimientos azarosos; es decir, no se debe escoger la muestra en función de las necesidades del investigador, sino de una manera completamente aleatoria. Otras cuestiones, como el tamaño de la muestra, también deberán ser tenidas en cuenta para la obtención de unos resultados representativos; basta considerar que una muestra de cien personas no será igual de representativa con respecto a una población de cien mil personas, que con respecto a una de diez millones.

Asimismo, la población no será siempre la misma aunque se considere a un mismo conjunto de individuos, pues dependerá del fenómeno que estemos estudiando. Observemos la diferencia que establecen Hernández Campoy y Almeida (2005, p.199):

> *Si queremos estudiar la madurez sintáctica de los escolares de la ciudad de Córdoba, la población estaría constituida por todos los individuos escolarizados de la ciudad, mientras que si queremos estudiar en la misma ciudad el uso de [l] por [r] (por ejemplo, en palabras como tarde/talde) en función de la edad, la clase social y la procedencia de los hablantes, la población la constituirían todos los habitantes de la ciudad que tengan un dominio aceptable del español.*

Lógicamente, hay alternativas al muestreo para estudiar una población, como el censo (observación exhaustiva, es decir, de todos y cada uno de los individuos de una población), la subpoblación (el estudio de una parte de la población que cumpla unas características determinadas), la observación mixta, etc. El motivo de elegir la muestra es porque resulta lo más aleatorio posible, si se elabora bien la selección muestral, como pone de manifiesto Casas Sánchez y Gutiérrez López (2011, p.15): "diremos que se ha investigado la población a partir de una muestra cuando los elementos que componen la muestra no reúnen ninguna característica esencial que los diferencie de los restantes, representando, por tanto, a toda la población", y porque la mayoría de los métodos alternativos, especialmente el censo, serían muy costosos de llevar a cabo.

También habrá que prestar atención a la actitud del investigador y a las características del informante. El investigador no debe persuadir o condicionar al informante, ni dejarse llevar por su posible conocimiento de la población estudiada, así como tampoco debe elegir informantes poco representativos, aunque formen parte de la población. Un informante poco representativo sería, por ejemplo, una persona que viaja con frecuencia a otras regiones y, por tanto, su manera de hablar no será tan representativa como la de una persona que nunca haya salido del lugar estudiado. Vemos una clarificación de esto en Paiva Boléo (1974, p.31):

> *A pessoa que fizer o inquérito não pode basear-se no conhecimento que porventura tenha do falar local, por mais profundo que ele seja, nem confiar nas informações de pessoas da classe ilustrada, por melhores conhecedores que sejam da linguagem da terra; deve interrogar um homem ou mulher do povo, que satisfaça, tanto quanto possível, aos seguintes requisitos: não ser desdentado; ter nascido e ter vivido sempre na terra; não se envergonhar da sua linguagem, mesmo diante de pessoas mais novas que já pronunciam as palavras doutra maneira (. . .); ter memória pronta; ser normalmente inteligente e, de preferência, analfabeto.*

En conclusión, recurriremos a la muestra por ser un método menos costoso al tiempo que representativo; sin embargo, para que esto último se cumpla, tendremos que ajustarnos a una selección adecuada de los informantes, sin que esto suponga desvirtuar la aleatoriedad que debe presentar la muestra. Es decir, la aleatoriedad no consiste en escoger a un individuo cualquiera independientemente de su representatividad, sino en elegir de manera aleatoria a los individuos de entre los potenciales informantes representativos de la población estudiada.

Medidas de posición central

La media

Según López Morán y Hernández Alonso (2012, p.44), "la media aritmética de un conjunto de datos es aquel valor que satisface la ecuación":

$$\bar{X} = \frac{\sum_{i=1}^{k} x_i n_i}{n}$$

Es decir, la media de un conjunto de datos será el sumatorio del producto de cada uno de los datos por el número de sus observaciones, dividido por el número total de observaciones.

Veamos cómo Hernández Campoy y Almeida (2005, p.204) ilustran una aplicación de la media aritmética al campo de la lingüística:

> *Si tras analizar una serie de datos sobre la altura del primer formante de [o] entre los hombres y las mujeres de una comunidad encontramos que los primeros registran una media de 465 Hz y las segundas una media de 492 Hz, podemos concluir que los hombres de dicha comunidad tienden a pronunciar la vocal de un modo más cerrado que las mujeres (dado que existe una relación entre la altura de F1 y la abertura vocálica: cuanto más alto es el valor de F1 más abierta será la vocal, y a la inversa).*

Basándonos en la cita previa, construiremos un ejemplo propio que resolveremos utilizando, como será habitual en todo el trabajo, R y/o Python -un lenguaje de programación especialmente diseñado para abordar cuestiones estadísticas, el primero; un potente lenguaje polivalente y de pleno auge en el terreno de las investigaciones científicas, el segundo-.

Ejemplo 1.- Supongamos un experimento en el que grabamos a 10 hombres y a 10 mujeres. Les pedimos que graben su voz pronunciando la vocal [o], los resultados que se obtienen son:

HOMBRES		MUJERES	
Hablante nº	Hercios	Hablante nº	Hercios
1	439	1	491
2	461	2	482
3	473	3	503
4	458	4	479
5	452	5	488
6	471	6	496
7	469	7	501
8	463	8	489
9	444	9	478
10	450	10	499

El tratamiento de la media en R es bastante simple; en primer lugar, asignamos a una variable el conjunto de los datos de los hombres y, a continuación, aplicamos la función para obtener la media implementada en R. Luego procedemos de igual manera con las mujeres:

```
>hombres = c(439,461,473,458,452,471,469,463,444,450)
>mean(hombres)
>[1] 458
>mujeres = c(491,482,503,479,488,496,501,489,478,499)
>mean(mujeres)
>[1] 490.6
```

Luego la media de realización de la [o] es de 458 Hz en el caso de los hombres y de 490.6 Hz en el caso de las mujeres. Esto quiere decir que las mujeres pronuncian esta vocal más abierta que los hombres.

Tras su resolución con R, procedemos a verificar que los resultados son idénticos en Python. Recomendamos que el usuario utilice el lenguaje con el que se sienta más cómodo. Lo primero que haremos será definir la función “media” que, a diferencia de en R, la escribiremos nosotros mismos en esta ocasión -esto es algo que aconsejamos en la medida de lo posible,

pues nos familiariza con los lenguajes de programación y nos servirá de base para enfrentarnos a problemas más complejos que requieran de cierta abstracción-.

```
def media(listadedatos):
    sumatorio = sum(listadedatos)
    númerodedatos = len(listadedatos)
    media = sumatorio/númerodedatos
    return media
```

Una vez definida la función, podremos utilizarla para obtener la media de cualquier conjunto de datos. Continuaremos con el mismo procedimiento utilizado en R:

```
>>>hombres = [439,461,473,458,452,471,469,463,444,450]
>>>media(hombres)
458.0
>>>mujeres = [491,482,503,479,488,496,501,489,478,499]
>>>media(mujeres)
490.6
```

Por tanto, se verifica que tanto la solución hallada con R como la hallada con Python son idénticas. La media obtenida puede ser solo hasta cierto punto representativa, pues si bien puede servir para orientarnos, no nos informa en absoluto de aspectos como la distribución de los datos, tanto en cuestiones sobre la manera en que están agrupados como la distancia que las observaciones presentan con respecto a la media obtenida, así como otras interpretaciones de interés. Por este y otros motivos, es necesario seguir investigando otras herramientas que irán supliendo estas carencias, así como conseguir perspectivas nuevas que nos harán sacar más provecho de los datos y, en consecuencia, enriquecer nuestras investigaciones sociolingüísticas, de cara a otorgarles un mayor rigor científico y sistemático.

La mediana

Según López Morán y Hernández Alonso (2012, p.45), la mediana "es aquel valor de la variable que divide la distribución de frecuencias en dos partes iguales. Es decir, aquel valor que deja a un lado y a otro el mismo número de observaciones."

El procedimiento será diferente dependiendo del tipo de distribución (valores sin agrupar o valores agrupados).

Sin agrupar

Buscamos primera frecuencia acumulada (N_i) que sea mayor o igual que la mitad del número total de observaciones ($\frac{N}{2}$).

$$N_i > \frac{N}{2} \rightarrow M_e = X_i$$

$$N_i = \frac{N}{2} \rightarrow M_e = \frac{X_i + X_{i+1}}{2}$$

Siendo N el número total de observaciones, N_i la frecuencia absoluta acumulada, X_i el dato o valor que estemos considerando, y M_e la mediana.

Agrupados

Buscamos primera frecuencia acumulada (N_i) que sea mayor o igual que la mitad del número total de observaciones ($\frac{N}{2}$) y aplicamos la siguiente fórmula:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i$$

Siendo L_{i-1} el valor correspondiente al extremo menor del intervalo considerado, y a_i su amplitud.

Veamos cómo Hernández Campoy y Almeida (2005, p.210) ilustran una aplicación de la mediana al campo de la lingüística:

Imaginemos una investigación en la que estudiamos ciertos cambios léxicos que se producen en una comunidad a partir de la incorporación de palabras foráneas y deseamos averiguar cuál es el grupo de edad que deja por encima y por debajo de él al 50% de la población, una información que puede resultar valiosa a la hora de determinar la fuerza con que ha penetrado y se ha instalado el cambio. Para ello se establece una clasificación de los informantes agrupados por grupos de edad donde se especifiquen tanto las frecuencias absolutas como las acumuladas.

A partir de la propuesta citada, crearemos una tabla cuyos datos estén dispuestos de la manera en que se nos sugiere y solucionaremos el ejercicio mediante el uso de Python (consideramos adecuada la alternancia entre diversos lenguajes de programación por si uno nos puede solucionar una tarea de manera más sencilla o eficiente, o por si simplemente el lector se siente más cómodo en alguno de ellos; sin embargo, se advertirá un mayor uso de Python, en lugar de R, pues es el lenguaje habitualmente usado en la investigación científica actual y es, por tanto, con el que consideramos que el usuario se sentirá más cómodo, a pesar de ser R un lenguaje específicamente pensado para el tratamiento estadístico de datos).

Ejemplo 2.- La tabla que hemos elaborado para este ejercicio es la siguiente:

Edad	Frecuencia	F. acumulada
16-25	19	19
26-35	17	36
36-45	16	52
46-55	11	63
56-65	8	71
66-75	4	75
76-85	1	76

A continuación procedemos a definir una función en Python que nos sirva para calcular la mediana en este tipo de casos. Para ello, lo primero que el lector debe hacer es asegurarse

de que comprende la definición de mediana, con el objetivo de poder abstraer el concepto y plasmarlo en código entendible por una máquina:

```
def mediana(frecuencias,intervalos):
    totalfrecuencias = sum(frecuencias)
    mitadfrecuencias = totalfrecuencias/2
    for i in range(len(frecuencias)):
        if sum(frecuencias[0:i+1]) ¿= mitadfrecuencias:
            intervalo = i
            break
    extremomenor = intervalos[intervalo][0]
    frecuenciaacumuladaanterior = sum(frecuencias[0:intervalo])
    frecuenciadelintervalo = frecuencias[intervalo]
    amplituddelintervalo = intervalos[intervalo][1]-intervalos[intervalo][0]
    mediana = extremomenor+(mitadfrecuencias-frecuenciaacumuladaanterior)/
frecuenciadelintervalo*amplituddelintervalo
    return mediana
```

Ahora, utilizando la función que acabamos de definir, hallaremos el valor de la mediana para el conjunto de datos considerado:

```
>>>mediana([19,17,16,11,8,4,1],[(16,25),(26,35),(36,45),(46,55),(56,65),(66,75),(76,85)])
37.125
```

Luego la mediana de la edad para estos datos agrupados sería 37.125 años. Esto quiere decir que el 50% de las observaciones se encuentran a la izquierda de 37.125 años, y el otro 50% a su derecha. Dicho de otro modo, si escogemos al azar a una persona de entre las entrevistadas, existirá la misma probabilidad de escoger a una persona menor de unos 37 años que de escoger a una persona mayor de tal edad.

La moda

Según López Morán y Hernández Alonso (2012, p.46), la moda "es el valor de la variable al que corresponde la máxima frecuencia. Es decir, el valor que más se repite en las observaciones."

El procedimiento será diferente dependiendo del tipo de distribución (valores sin agrupar o valores agrupados).

Sin agrupar

La moda será, sencillamente, el valor con la frecuencia absoluta más alta; es decir, el valor que más se repite en términos absolutos.

Agrupados

A su vez, distinguimos dos procedimientos dentro de los agrupados, atendiendo a la amplitud de los intervalos (amplitud constante o amplitud no constante).

Amplitud constante $\rightarrow$ Encontramos la n_i mayor y aplicamos la siguiente fórmula:

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i+1}+n_{i-1}} a_i$$

Amplitud no constante $\rightarrow$ Encontramos la h_i mayor y aplicamos la siguiente fórmula:

$$M_o = L_{i-1} + \frac{h_{i+1}}{h_{i+1}+h_{i-1}} a_i$$

Siendo cada h_i cada uno de los respectivos cocientes $\frac{n_i}{a_i}$.

Veamos una aplicación de la moda propuesta por Hernández Campoy y Almeida (2005, p.209): "Supongamos (...) un test sobre actitudes hacia una variedad dialectal, en el que se hace escuchar dos voces a una serie de informantes con el fin de que estos las evalúen como "agradable", "desagradable" e "indiferente"."

Basándonos en su propuesta, vamos a idear un ejemplo al que daremos solución utilizando tanto R como Python.

Ejemplo 3.- Entrevistamos a 10 personas y les pedimos que clasifiquen a dos voces en base a las categorías antes mencionadas. Los datos obtenidos se tabulan a continuación:

Participante	Voz 1	Voz 2
1	Agradable	Indiferente
2	Indiferente	Indiferente
3	Agradable	Agradable
4	Indiferente	Indiferente
5	Agradable	Indiferente
6	Desagradable	Indiferente
7	Indiferente	Desagradable
8	Agradable	Indiferente
9	Desagradable	Indiferente
10	Agradable	Agradable

Comenzaremos con R, y para ello escribimos las siguientes instrucciones:

>voz1 = c('agradable', 'indiferente', 'agradable', 'indiferente', 'agradable', 'desagradable', 'indiferente', 'agradable', 'desagradable', 'agradable')

>voz2 = c('indiferente', 'indiferente', 'agradable', 'indiferente', 'indiferente', 'indiferente', 'desagradable', 'indiferente', 'indiferente', 'agradable')

>table(voz1)

voz1

agradable desagradable indiferente
5 2 3

>table(voz2)

voz2

agradable desagradable indiferente
2 1 7

```
>which.max(table(voz1))
agradable
1

>which.max(table(voz2))
indiferente
3
```

En primer lugar, hemos asignado a las variables *voz1* y *voz2* los datos correspondientes del ejercicio. A continuación, le pedimos que nos muestre una sencilla tabla en la que se nos indica la frecuencia de cada uno de estos atributos -lo hacemos con un carácter meramente informativo-; por último, le pedimos que nos devuelva el más frecuente de los atributos -junto a su posición en la tabla construida-. Hemos obtenido que la moda para *voz1* es 'agradable', mientras que la moda para *voz2* es 'indiferente'.

Ahora procedemos a su resolución con Python:

```
def moda(datos):
    diccionario = {}
    for i in set(datos):
        diccionario[i] = datos.count(i)
    ordenado = sorted(diccionario, key=diccionario.get, reverse=True)
    return ordenado[0]

>>>voz1 = ['agradable', 'indiferente', 'agradable', 'indiferente', 'agradable', 'desagradable',
'indiferente', 'agradable', 'desagradable', 'agradable']
>>>voz2 = ['indiferente', 'indiferente', 'agradable', 'indiferente', 'indiferente', 'indiferente',
'desagradable', 'indiferente', 'indiferente', 'agradable']
>>>moda(voz1)
'agradable'
>>>moda(voz2)
'indiferente'
```

Corroboramos lo que ya obtuvimos con R; que la moda de *voz1* es 'agradable' y de *voz2* 'indiferente'.

Medidas de dispersión absolutas

La varianza

Según López Morán y Hernández Alonso (2012, p.53), la varianza "es la media aritmética de los cuadrados de las desviaciones de los valores de la variable a la media aritmética:"

$$S^2 = \frac{\sum_{i=1}^{k} (X_i - \bar{X})^2 n_i}{n}$$

El inconveniente que presenta la varianza es que resulta poco intuitiva por venir dada en unidades cuadradas. Por ello, se recurre a la desviación típica.

La desviación típica

La desviación típica es el resultado de aplicarle la raíz cuadrada a la varianza. De esta manera, obtenemos una unidad más fácilmente interpretable y de gran utilidad para hacernos una idea de si la dispersión de un conjunto de datos respecto a su media es muy alta o, por el contrario, los datos se encuentran relativamente próximos a la media del conjunto.

$$S = \sqrt{S^2}$$

Para una ampliación de la importancia de la varianza y la desviación típica como representantes de la dispersión de los datos, consúltese en Martín-Pliego López y Ruiz-Maya Pérez (2010, pp.137-141) y en Casas Sánchez y Gutiérrez López (2011, pp.161-162) la **Desigualdad de Chebychev**, así como una simplificación del **Teorema de Chebychev** en López Morán y Hernández Alonso (2012, p.54), que no detallamos aquí por no considerarlo imprescindible dentro de los límites que abarca esta guía básica.

Examinemos la propuesta de Hernández Campoy y Almeida (2005, pp.216-217) para entender mejor una aplicación de la varianza y la desviación típica a un problema sociolingüístico:

Supongamos que en una investigación deseamos estudiar la incidencia que tiene la interferencia del inglés en una comunidad hispana a través del análisis de una serie de préstamos integrados o aceptados (o parcialmente integrados) del tipo "formatear". Deseamos saber, además, si existen diferencias entre hombres y mujeres, ya que debido al tipo de actividad laboral que desempeñan los miembros de cada grupo en dicha comunidad se esperaba que ambos grupos tuvieran comportamientos discrepantes. Para ello se analizan dos horas de conversación con cada informante y se anota para cada individuo el número de unidades léxicas emitidas.

Ejemplo 4.- Basándonos en lo anterior, llevamos a cabo un experimento hipotético en el que obtenemos los siguientes resultados. A ellos, les aplicaremos la varianza y la desviación típica y, por último, estableceremos las conclusiones pertinentes.

Hombre nº	Emisiones	Mujer nº	Emisiones
1	5	1	3
2	3	2	1
3	4	3	6
4	4	4	2
5	6	5	5
6	3	6	2
7	5	7	6
8	3	8	3
9	5	9	5
10	6	10	1

Procedemos a su resolución con R:

```
>hombres = c(5,3,4,4,6,3,5,3,5,6)
>mujeres = c(3,1,6,2,5,2,6,3,5,1)
>varH = var(hombres)
>varH
>[1] 1.377778
>desH = sqrt(varH)
```

>desH

>[1] 1.173788

>varM = var(mujeres)

>varM

>[1] 3.822222

>desM = sqrt(varM)

>desM

>[1] 1.95505

Luego varH = 1.377778, desH = 1.173788, varM = 3.822222 y desM = 1.95505. De aquí, nos quedaremos con los dos datos más representativos:

Desviación típica en los hombres = 1.173788

Desviación típica en las mujeres = 1.95505

Esto quiere decir que los hombres presentan unos resultados más homogéneos, puesto que las observaciones se sitúan cerca de la media del conjunto; por el contrario, las mujeres presentan unos resultados más heterogéneos, ya que las observaciones -tal y como demuestra su mayor desviación típica- están más dispersas con respecto a la media de su conjunto.

Resolución con Python:

```
import math
def media(listadedatos):
    sumatorio = sum(listadedatos)
    númerodedatos = len(listadedatos)
    media = sumatorio/númerodedatos
    return media
def var(datos):
    mediadatos = media(datos)
    varianza = sum((i-mediadatos)**2 for i in datos)/len(datos)
    return varianza
def destip(datos):
    destip = math.sqrt(var(datos))
    return destip
```

Por último, aplicaremos las funciones a los datos del ejercicio, para comprobar los resultados:

```
>>>hombres = (5,3,4,4,6,3,5,3,5,6)
>>>var(hombres)
1.2399999999999998
>>>destip(hombres)
1.1135528725660042
>>>mujeres = (3,1,6,2,5,2,6,3,5,1)
>>>var(mujeres)
3.44
>>>destip(mujeres)
1.8547236990991407
```

Luego var(hombres) = 1.2399999999999998, destip(hombres) = 1.1135528725660042, var(mujeres) = 3.44 y des(mujeres) = 1.8547236990991407. De aquí, nos quedaremos con los dos datos más representativos:

Desviación típica en los hombres = 1.1135528725660042

Desviación típica en las mujeres = 1.8547236990991407

Los resultados tendrían una interpretación idéntica a la llevada a cabo anteriormente, en la resolución del ejercicio mediante R. Sin embargo, las varianzas y desviaciones típicas no coinciden totalmente entre los resultados arrojados por R y los resultados obtenidos en Python; esto se debe a que R utiliza un criterio ligeramente diferente para obtener la varianza -y, por tanto, la desviación típica-: “The denominator n - 1 is used which gives an unbiased estimator of the (co)variance for i.i.d. observations.” Esta cita y su contexto pueden ser encontrados en la documentación de R, apartado Correlation, Variance and Covariance (Matrices), en el caso de que se pretenda investigar el funcionamiento interno de la función implementada en R.

Llegados a este punto, nos gustaría llamar la atención sobre la importancia de la definición de funciones, frente a la inserción de datos directamente en la terminal o consola, no solo por el hecho de que nos ahorra tener que escribir lo mismo para todas las operaciones del mismo tipo, sino que, como hemos visto, las funciones pueden ser reutilizadas en el código de otra función, como hemos hecho con la definición de la media, que hemos usado en el código destinado a obtener la varianza, o la propia función de varianza, que nos ha resuelto de manera casi inmediata la función de la desviación típica.

Análisis de dos variables

Covarianza

Según López Morán y Hernández Alonso (2012, p.91), definimos la covarianza como "la media aritmética de los productos entre las desviaciones de la variable X respecto a su media y las de la variable Y respecto a la suya":

$$S_{XY} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{h} (X_i - \bar{X})(Y_j - \bar{Y}) n_{ij}$$

Es decir, la covarianza nos indica la interrelación existente entre dos caracteres; en qué manera uno sube si el otro también lo hace, o cómo uno sube mientras el otro baja, o incluso si sencillamente no habría cambios signficativos. Concretamente, la relación lineal entre las dos variables sería nula si la covarianza es nula; sin embargo, cuanto mayor, en términos absolutos, -por tanto, siendo distinto de cero- sea la covarianza, mayor será el grado de dependencia lineal. Recordemos que un signo positivo significa que la relación lineal que se establece es directa, mientras que si el signo de la covarianza es negativo, la relación lineal establecida será inversa. En cualquier caso, volveremos sobre este asunto cuando estudiemos más adelante el coeficiente de correlación lineal, así como la consecuencia que la generalización de la Desigualdad de Schwarz tiene en este sentido.

Observamos una propuesta de Hernández Campoy y Almeida (2005, p.231), en la que se nos dice: "Supongamos que deseamos estudiar si la variable Red social influye sobre (o convaría con) la variable Duración cuando analizamos las diferencias temporales de (p, t, k) entre hombres y mujeres (variable Sexo)"; como vemos, no se trata sino de verificar si existe alguna relación de dependencia entre una variable y otra.

Mientras que el enunciado anterior es ilustrativo para entender qué estamos buscando, no es un buen ejemplo de índole sociolingüístico, a pesar de estar contenido en un manual orientado a tal fin; por tanto, seremos nosotros los que propondremos un ejercicio al que daremos solución por los procedimientos habituales.

Ejemplo 5.- Supongamos un experimento en el que sometemos a 10 personas a dos contextos distintos: uno en el que el registro esperado es el coloquial, y otro en el que el registro esperado es el formal. En ambas ocasiones se les graba durante 60 minutos y luego confeccionamos una tabla en la que se recoja cuántas veces el individuo considerado ha proferido la palabra comodín "cosa", distinguiendo entre las veces que lo hizo en el contexto informal y las que lo hizo en el contexto formal.

Los datos obtenidos tras el hipotético experimento son:

Contexto informal	Contexto formal
13	3
15	7
12	4
8	1
11	5
17	8
19	7
15	5
12	6
10	2

Definiremos, pues, una función para hallar la covarianza. Posteriormente, la aplicaremos a los datos concretos de nuestro ejercicio. Para el cálculo de la covarianza, nos hemos basado en esta expresión, más simple que la de la definición:

$$S_{XY} = \frac{\sum n_i x_i y_i}{N} - \bar{X}\bar{Y}$$

```
def media(listadedatos):
    sumatorio = sum(listadedatos)
    númerodedatos = len(listadedatos)
    media = sumatorio/númerodedatos
    return media
```

```
def covar(listadedatos1,listadedatos2):
    media1 = media(listadedatos1)
    media2 = media(listadedatos2)
    productodemedias = media1*media2
    sumatorio = sum([listadedatos1[i]*listadedatos2[j] for i in
range(len(listadedatos1)) for j in range(len(listadedatos2)) if i == j])
    N = len(listadedatos1)
    covar = sumatorio/N - productodemedias
    return covar


>>>covar([13,15,12,8,11,17,19,15,12,10],[3,7,4,1,5,8,7,5,6,2])
5.740000000000002
```

Ahora, veamos cómo podemos abordar con R esta cuestión:

```
>coloquial = c(13,15,12,8,11,17,19,15,12,10)
>formal = c(3,7,4,1,5,8,7,5,6,2)
>cov(coloquial,formal)
6.377778
```

Como vemos, el resultado no es idéntico. Esto se explica por cómo R define la covarianza: $S_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})$

Una gran asociación demostraría que existe un patrón a la hora de emplear cierto léxico del tipo comodín, que consistiría en un mayor uso de estos en contextos informales, frente a una disminución de su uso en ambientes o contextos formales.

Para obtener una medida más intuitiva del grado de correlación entre ambas variables, nos serviremos, a partir de la covarianza, del coeficiente de correlación lineal, que acota el restultado entre -1 y 1.

Coeficiente de correlación lineal

Conforme a Martín-Pliego López y Ruiz-Maya Pérez (2010, p.149), "el coeficiente de correlación lineal mide el grado de asociación lineal entre las variables aleatorias ξ y η, y se define como":

$$\rho = \frac{cov(\xi;\eta)}{\sigma_\xi \sigma_\eta}$$

Como indica López Morán y Hernández Alonso (2012, p.104), "El valor del mismo se encuentra acotado, en términos absolutos, entre el valor cero y el valor uno. El valor uno nos indica covariación perfecta entre las variables y el valor cero covariación nula". Obviamente, los valores intermedios nos irán informando de la mayor o menos asociación entre las variables en función de si se aproximan o alejan del valor uno. Asimismo, deberemos prestar atención al signo del coeficiente, pues un signo positivo nos indicará una relación directa, mientras que uno negativo, inversa; en otras palabras, si el signo del coeficiente es positivo, querrá decir que según aumenta el valor de una variable, también lo hará el de la otra, mientras que si es negativo, el valor de una variable disminuirá al aumentar el de la otra variable. Concluimos, pues, que si las variables son independientes, el coeficiente de correlación lineal es nulo; sin embargo, lo contrario no es necesariamente cierto, pues dos variables aleatorias podrían estar aleatoriamente incorrelacionadas pero no ser estadísticamente independientes. Para una mayor comprensión de este punto, que puede resultar confuso, aconsejamos la lectura del ejemplo proporcionado en Martín-Pliego López y Ruiz-Maya Pérez (2010, pp.150-151).

Vamos a recurrir ahora a la ejemplificación, como es habitual, y para ello vamos a atender a lo propuesto por Hernández Campoy y Almeida (2005, p.239): "Supongamos que queremos estudiar la relación entre el número de formas lingüísticas vernáculas que emplean los individuos y su edad". Basándonos en el enunciado, proponemos el siguiente ejemplo:

Ejemplo 6.- Tras llevar a cabo un estudio en un pueblo de la sierra de Cádiz sobre el número de emisiones de formas vernáculas durante una grabación de sesenta minutos, atendiendo a la edad del informante, obtenemos la siguiente tabla:

Número de emisiones	Edad del informante
2	18
4	21
4	23
5	29
8	32
7	35
11	41
12	48
16	53
21	62

Procedemos a su resolución mediante una definición de función en Python, aprovechando las que ya tenemos previamente definidas, y su posterior aplicación a los datos concretos del ejemplo:

```
import math
def media(listadedatos):
    sumatorio = sum(listadedatos)
    númerodedatos = len(listadedatos)
    media = sumatorio/númerodedatos
    return media
def var(datos):
    mediadatos = media(datos)
    varianza = sum((i-mediadatos)**2 for i in datos)/len(datos)
    return varianza
def destip(datos):
    destip = math.sqrt(var(datos))
    return destip
def covar(listadedatos1,listadedatos2):
    media1 = media(listadedatos1)
    media2 = media(listadedatos2)
```

```
    productodemedias = media1*media2
    sumatorio = sum([listadedatos1[i]*listadedatos2[j] for i in
range(len(listadedatos1)) for j in range(len(listadedatos2)) if i == j])
    N = len(listadedatos1)
    covar = sumatorio/N - productodemedias
    return covar
def coefcorr(datos1,datos2):
    coefcorr = covar(datos1,datos2)/(destip(datos1)*destip(datos2))
    return coefcorr
```

Aplicamos ahora esta función que hemos definido a los datos de nuestro ejemplo:

```
>>>emisiones = (2,4,4,5,8,7,11,12,16,21)
>>>edades = (18,21,23,29,32,35,41,48,53,62)
>>>coefcorr(emisiones,edades)
0.9827277240659655
```

Verifiquemos el resultado que obtendríamos usando R:

```
>emisiones = c(2,4,4,5,8,7,11,12,16,21)
>edades = c(18,21,23,29,32,35,41,48,53,62)
>cor(emisiones,edades)
0.9827277
```

Como era de esperar, pues a simple vista se aprecia la gran asociación entre las dos variables, el coeficiente de correlación es, aproximadamente, 0.98; es decir, un valor muy próximo a 1, lo que confirma que a mayor edad, cabe esperar un mayor número de emisiones de formas vernáculas. Sabemos, además, que la relación es directa, y no inversa, gracias al signo positivo del coeficiente.

A pesar de que este coeficiente, como acabamos de comprobar, resulta de gran utilidad para comprobar la intensidad de la asociación entre dos variables, no nos permite calcular el valor exacto que, dado el valor de una variable, tendría la otra. Para ello, vamos a recurrir a la regresión lineal.

Regresión lineal y bondad del ajuste

Siguiendo a López Morán y Hernández Alonso (2012, p.105), "con esta técnica se busca, expresamente, determinar una función matemática (f), denominada ecuación de regresión, que refleje, del modo más exacto posible, la relación existente entre las variables X e Y". Buscamos, pues, una expresión matemática del tipo $y = f(x)$ que explique el comportamiento de una variable en función de la otra; en nuestro caso, lo que buscamos es determinar una recta que se aproxime lo máximo posible a las observaciones dadas para, a partir de ellas, obtener una función general que permita describir el fenómeno ante observaciones no consideradas y proporcionarnos un valor aproximado de una variable, dado el valor de la otra. Por tratarse de una recta, la función que buscamos tendrá esta apariencia:

$$Y = a + bX$$

Tal y como indican Martín-Pliego López y Ruiz-Maya Pérez (2010, p.265), "se utiliza el principio de los mínimos cuadrados, de forma que la función (...) que buscamos debe verificar que el cuadrado del error cometido en la regresión de η respecto a ξ sea mínimo". Dicho de otra manera, buscaremos que la recta de regresión se aleje lo menos posible de las observaciones dadas. Para un estudio más detallado de todo el proceso y fundamentación de la obtención de la recta de regresión lineal, aconsejamos la lectura de Martín-Pliego López y Ruiz-Maya Pérez (2010, pp.259-267).

A nuestros efectos, nos quedaremos con una visión más funcional del concepto, y nos limitaremos a explicar, de manera sencilla, cómo obtener tal recta de regresión. Teniendo en cuenta la ecuación de la recta antes descrita, resulta obvio que lo que buscamos son los valores de a y de b, pues X es, precisamente, nuestro dato a partir del cual hallaremos el valor de Y. Gracias a estas expresiones, podremos encontrar los valores de a y de b fácilmente:

$$b = \frac{S_{XY}}{S_X^2} \; ; \; a = \bar{Y} - b\bar{X}$$

Siendo "a" el término independiente (ordenada en el origen), y "b" la pendiente de la recta (cuánto se incrementa Y en relación a X).

Basándonos en la misma idea que en el ejemplo 6, proponemos el siguiente:

Ejemplo 7.- En un experimento, estudiamos el número de emisiones de arcaísmos durante una conversación grabada de sesenta minutos, en un pueblo determinado. Además de las emisiones, se hará constar la edad de cada uno de los informantes, de cara a intentar determinar una recta de regresión lineal que explique dicho comportamiento en el pueblo estudiado. Se obtienen los siguientes resultados:

Número de emisiones	Edad del informante
1	18
1	20
2	21
1	28
4	33
5	39
10	45
12	48
13	56
17	63

Definiremos una función en Python que nos ayude a calcular la recta que buscamos, aprovechando para ello otras funciones previamente definidas en ejercicios anteriores:

```
import math
def media(listadedatos):
    sumatorio = sum(listadedatos)
    númerodedatos = len(listadedatos)
    media = sumatorio/númerodedatos
    return media
def covar(listadedatos1,listadedatos2):
    media1 = media(listadedatos1)
    media2 = media(listadedatos2)
    productodemedias = media1*media2
```

```
    sumatorio = sum([listadedatos1[i]*listadedatos2[j] for i in
range(len(listadedatos1)) for j in range(len(listadedatos2)) if i == j])
    N = len(listadedatos1)
    covar = sumatorio/N - productodemedias
    return covar
def var(datos):
    mediadatos = media(datos)
    varianza = sum((i-mediadatos)**2 for i in datos)/len(datos)
    return varianza
def reglineal(datos1,datos2):
    b = covar(datos1,datos2)/var(datos1)
    a = media(datos2)-(b*media(datos1))
    return "Y = "+"%f "%a+"+ "+"%f"%b+"X"
```

Mostramos a continuación la aplicación de la función definida a los datos del ejemplo:

```
>>>edades = (18,20,21,28,33,39,45,48,56,63)
>>>emisiones = (1,1,2,1,4,5,10,12,13,17)
>>>reglineal(edades,emisiones)
"Y = -6.905738 + 0.364036X"
```

Veamos cómo plantearlo con R:

```
>edades = c(18,20,21,28,33,39,45,48,56,63)
>emisiones = c(1,1,2,1,4,5,10,12,13,17)
>lm(emisiones  edades)
Call:
lm(formula = emisiones  edades)

Coefficients:
(Intercept)        edades
-6.906             0.364
```

Para comprobar que el resultado sea razonable, podemos probar a sustituir la X por uno de los valores observados y comprobar que su correspondiente Y se aproxime al valor correspondiente observado. Si tomamos, por ejemplo, X = 45, obtenemos Y = 9.475882 (siendo 10 el original observado), por lo que confirmamos que nuestra recta de regresión lineal está, en principio, bien ajustada.

A pesar de que lo que acabamos de decir no está errado, por lo general, es preciso recurrir a algún tipo de medición precisa de la bondad del ajuste, pues no siempre estará tan claro si nuestro ajuste ha sido el esperado, o no. Para ello, vamos a recurrir al coeficiente de determinación, que es una medida de dispersión relativa y viene definida por la expresión, acotada entre 0 y 1:

$$R^2 = \frac{(S_{XY})^2}{S_X^2 S_Y^2}$$

Como indican López Morán y Hernández Alonso (2012, p.110), "El valor cero debe tomarse como indicativo de nula representatividad de la ecuación ajustada (covariación nula) y el valor uno como expresión de un ajuste perfecto o concordancia plena entre recta de regresión y observaciones (covariación exacta)".

Definamos en Python una función para obtener R^2 y obtengamos su valor para las observaciones del último ejemplo:

```
import math
    def media(listadedatos):
    sumatorio = sum(listadedatos)
    númerodedatos = len(listadedatos)
    media = sumatorio/númerodedatos
    return media
def covar(listadedatos1,listadedatos2):
    media1 = media(listadedatos1)
    media2 = media(listadedatos2)
    productodemedias = media1*media2
    sumatorio = sum([listadedatos1[i]*listadedatos2[j] for i in
```

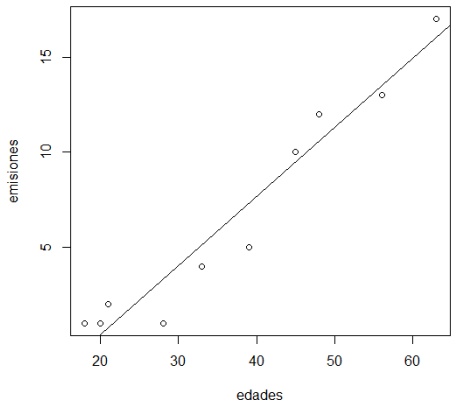
```
range(len(listadedatos1)) for j in range(len(listadedatos2)) if i == j])
        N = len(listadedatos1)
        covar = sumatorio/N - productodemedias
        return covar
def var(datos):
        mediadatos = media(datos)
        varianza = sum((i-mediadatos)**2 for i in datos)/len(datos)
        return varianza
def coefdet(datos1,datos2):
        coefdet = (covar(datos1,datos2)**2)/(var(datos1)*var(datos2))
        return coefdet
```

Y su aplicación:

```
>>>edades = (18,20,21,28,33,39,45,48,56,63)
>>>emisiones = (1,1,2,1,4,5,10,12,13,17)
>>>coefdet(edades,emisiones)
0.9395002178758177
```

El valor de coeficiente de determinación es realmente alto, pues $R^2 = 0{,}94$ aproximadamente; es decir, un valor muy próximo a 1, lo que demuestra el alto grado de bondad de nuestro ajuste.

Por último, presentamos un gráfico construido en R en el que se muestra la nube de puntos (observaciones), así como la recta de regresión:



Prueba estadística básica

Chi cuadrado χ^2

La prueba χ^2 se basa en que, tal y como señalan Casas Sánchez y Gutiérrez López (2011, p.304), "si la hipótesis nula H_0 es cierta, las frecuencias observadas n_i no deberían desviarse mucho de sus valores esperados np_i (i = 1,..., k). Luego parece lógico utilizar, como estadístico de prueba, un estadístico que nos mida las discrepancias entre ambos valores". Por tanto, podríamos definirla así:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Esta prueba, que es un medidor de la bondad de ajuste, es ampliamente utilizada para su aplicación a conjuntos de datos nominales, especialmente en el ámbito de la lingüística en general, y de la sociolingüística en particular. Veamos una posible aplicación propuesta por Hernández Campoy y Almeida (2005, 233):

> *Supongamos que deseamos estudiar si existen diferencias en el uso de eufemismos en una comunidad en relación con la clase social de los individuos. De acuerdo a la tesis, ampliamente corroborada, de que las clases altas suelen emplear las formas lingüísticas que se valoran prestigiosas con más frecuencia que las demás clases, se esperaba que los usos lingüísticos reprodujeran fielmente la estratificación social, de modo que los valores más altos en el uso de eufemismos se encontrara en la clase alta y los más bajos en la clase baja.*

Basándonos en la idea propuesta, presentamos el siguiente ejemplo:

Ejemplo 8.- Tras estudiar, en el contexto de una investigación sociolingüística, el número de emisiones de formas lingüísticas prestigiadas atendiendo al estrato social de los informantes, durante una conversación grabada de sesenta minutos, se obtienen los resultados que exponemos a continuación:

Estrato del informante	Número de emisiones
Bajo	2
Bajo	1
Bajo	4
Medio	6
Medio	9
Medio	8
Alto	16
Alto	13
Alto	17

Definamos una sencilla función en Python que nos permita calcular el valor de χ^2:

```
def chicuad(datosobservados,datoteórico):
    sumatorio = sum(((i[1]-datoteórico)**2)/datoteórico for i in observaciones)
    return sumatorio
```

Ahora, apliquemos la función a los datos del ejemplo propuesto:

```
>>>observaciones = (('bajo',2),('bajo',1),('bajo',4),('medio',6),('medio',9),('medio',8),
('alto',16),('alto',13),('alto',17))
>>>observaciónteórica = 7
>>>chicuad(observaciones,observaciónteórica)
41.85714285714286
```

Puesto que el resultado se aleja de 0, sabemos que habrá un grado de asociación entre las variables, pues este grado será mayor en tanto que nos alejemos del valor nulo; el inconveniente, sin embargo, reside en que χ^2 no está acotado superiormente y, por tanto, no podemos calibrar ese grado de asociación; es decir, dada la χ^2, no estaremos seguros de si el grado de asociación es bajo, medio, alto o muy alto, a pesar de ser distinta de cero. Para solucionar este problema, recurrimos al coeficiente de contingencia de Pearson.

Coeficiente de contingencia de Pearson

Este coeficente, conforme señalan López Morán y Hernández Alonso (2012, p.153), "tiene la ventaja de estar acotado entre cero y uno. El cero indica (...) carencia de asociación y el uno, asociación perfecta". La expresión que define este coeficiente es:

$$C = \sqrt{\frac{\chi^2}{N+\chi^2}}$$

Para ver su aplicación, lo utilizaremos para clarificar el resultado de χ^2 obtenido en el ejemplo anterior. Para ello, comenzaremos definiendo una función en Python que nos generalice el método de obtención del coeficiente:

```
import math
def chicuad(datosobservados,datoteórico):
    sumatorio = sum(((i[1]-datoteórico)**2)/datoteórico for i in observaciones)
    return sumatorio
def coefcontpearson(númobs,chi):
    coef = math.sqrt(chi/(númobs+chi))
    return coef
```

Y la aplicamos a los datos y resultados del ejemplo anterior:

```
>>>observaciones = (('bajo',2),('bajo',1),('bajo',4),('medio',6),('medio',9),('medio',8),
('alto',16),('alto',13),('alto',17))
>>>observaciónteórica = 7
>>>coefcontpearson(len(observaciones),chicuad(observaciones,observaciónteórica))
0.9072120523147654
```

Comprobamos que, como esperábamos, el resultado 0.9 se encuentra muy próximo a 1, que sería la asociación teórica perfecta. Podemos concluir, pues, sin miedo a equivocarnos, que existe un alto grado de asociación entre la variable clase social y la variable emisiones de formas lingüísticas prestigiadas en la población estudiada.

Ejercicio propuesto

Ejercicio.- Supongamos un experimento realizado sobre la población de un municipio, en el que intentemos verificar la relación, o no, entre la edad de los sujetos y el número de emisiones de vulgarismos. Tras realizar grabaciones de 60 minutos sobre los sujetos, se obtienen los siguientes resultados:

Edad del informante	Número de emisiones
19	16
25	14
29	15
34	12
38	9
43	7
47	7
56	3
61	2
68	1

A partir de los datos del ejercicio, obtener y dar una interpretación sociolingüística a:

a) La media de las emisiones.
b) La moda de las emisiones.
c) La desviación típica de las emisiones.
d) El coeficiente de correlación lineal entre las variables.
e) La recta de regresión lineal.
f) El coeficiente de determinación.
g) El coeficiente de contingencia de Pearson.

Recordamos que, para abordar convenientemente el ejercicio, se tendrá que plantear teóricamente, en primer lugar, para luego darle un tratamiento computacional y, por último, interpretar a nivel sociolingüístico los resultados obtenidos.

Ejercicio propuesto resuelto

En primer lugar, tendremos que definir en Python todas las funciones que utilizaremos a continuación, y que nos proporcionarán, de manera inmediata, los resultados que buscamos. Obviamente, si ya los hemos definido en ejercicios previos, podremos usarlos directamente, lo cual agiliza, aún más, nuestra tarea:

```
import math
def media(listadedatos):
    sumatorio = sum(listadedatos)
    númerodedatos = len(listadedatos)
    media = sumatorio/númerodedatos
    return media
def moda(datos):
    diccionario =
    for i in set(datos):
        diccionario[i] = datos.count(i)
    ordenado = sorted(diccionario, key=diccionario.get, reverse=True)
    return ordenado[0]
def var(datos):
    mediadatos = media(datos)
    varianza = sum((i-mediadatos)**2 for i in datos)/len(datos)
    return varianza
def destip(datos):
    destip = math.sqrt(var(datos))
    return destip
def covar(listadedatos1,listadedatos2):
    media1 = media(listadedatos1)
    media2 = media(listadedatos2)
    productodemedias = media1*media2
    sumatorio = sum([listadedatos1[i]*listadedatos2[j] for i in range(len(listadedatos1))
for j in range(len(listadedatos2)) if i == j])
```

```
    N = len(listadedatos1)
    covar = sumatorio/N - productodemedias
    return covar
def coefcorr(datos1,datos2):
    coefcorr = covar(datos1,datos2)/(destip(datos1)*destip(datos2))
    return coefcorr
def reglineal(datos1,datos2):
    b = covar(datos1,datos2)/var(datos1)
    a = media(datos2)-(b*media(datos1))
    return "Y = -" %f " %a+- -" %f" %b+"X"
def coefdet(datos1,datos2):
    coefdet = (covar(datos1,datos2)**2)/(var(datos1)*var(datos2))
    return coefdet
def chicuad(datosobservados,datoteórico):
    sumatorio = sum(((i[1]-datoteórico)**2)/datoteórico for i in observaciones)
    return sumatorio
def coefcontpearson(númobs,chi):
    coef = math.sqrt(chi/(númobs+chi))
    return coef
```

Ahora, utilizaremos estas funciones para aplicárselas a los datos del ejercicio y así poder interpretar posteriormente los resultados:

```
>>>edades = (19,25,29,34,38,43,47,56,61,68)
>>>emisiones = (16,14,15,12,9,7,7,3,2,1)
>>>media(emisiones)
8.6
>>>moda(emisiones)
7
>>>destip(emisiones)
5.238320341483519
>>>coefcorr(edades,emisiones)
-0.9825909233381221
```

```
>>>reglineal(edades,emisiones)
'Y = 22.774549 + -0.337489X'
>>>coefdet(edades, emisiones)
0.9654849226264636
>>>observaciones = [(edades[i],emisiones[i]) for i in range(len(edades))]
>>>teórica = 10
>>>coefcontpearson(len(observaciones),chicuad(observaciones,teórica))
0.8638245732792135
```

En resumen, los datos obtenidos son:

Media de las emisiones: 8.6

Moda de las emisiones: 7

Desviación típica de las emisiones: 5.24

Coeficiente de correlación lineal: -0.98

Recta de regresión lineal: y = 22.774549 + -0.337489x

Coeficiente de determinación: 0.97

Coeficiente de contingencia de Pearson: 0.86 (con frecuencia teórica = 10)

Las conclusiones son, pues, que los datos se distribuyen de una manera poco concentrada (obsérvese que la desviación típica es muy alta en relación a la media), que el grado de asociación entre las variables es altísimo, pero de sentido inverso, por tener el coeficiente de correlación lineal signo negativo (esto quiere decir que a medida que aumenta el valor de una variable, disminuirá la otra), que al ser el coeficiente de determinación tan alto podemos tener una gran seguridad en relación a la bondad de nuestro ajuste de relación lineal y, por tanto, podremos tener una noción bastante aproximada de la relación entre valores no observados y, finalmente, tenemos la certeza, de nuevo, de que el grado de asociación es alto, gracias al valor próximo a 1 que presenta el coeficiente de contigencia de Pearson.

A partir de la clarificación del sentido de los resultados, podríamos afirmar, con poco miedo a equivocarnos, que la edad de los informantes es un factor decisivo con respecto al número de emisiones de vulgarismos que habitualmente emite y, gracias a la recta de regresión lineal elaborada, podríamos aproximar el número de vulgarismos medio que una persona de una edad determinada emitiría, en la población estudiada.

Conclusiones

Tras el recorrido práctico -con pinceladas de apuntes teóricos- de las herramientas estadísticas básicas al alcance de los investigadores en sociolingüística, pretendemos haber clarificado cuestiones como la necesidad de incorporación de formación estadística en el currículo formativo de los alumnos e investigadores en sociolingüística, y la oportunidad y viabilidad de la incorporación de dicha formación, siempre que el enfoque no sea convertir a los lingüistas en estadísticos, sino en investigadores autónomos que puedan realizar sus investigaciones, de manera sistemática y científica, apoyados en sólidas teorías matemáticas, sin la necesidad de un colaborador externo que supla las carencias del manejo estadístico de datos.

Nos gustaría, también, resaltar la importancia de servirse de herramientas computacionales en el uso de la estadística, pues facilita el trabajo del investigador, además de allanar el camino para futuras investigaciones, con la creación o uso de expresiones o funciones predefinidas y dar cuenta de la diferencia entre usar herramientas usadas por otros (nuestro uso habitual de R), o la gran ventaja de crear nuestras propias funciones (como hemos procedido en Python), porque nos abre la puerta a personalizar las funciones hasta el punto de que nos proporcionen exactamente lo que buscamos, cuestión más complicada cuando usamos funciones o programas de terceros.

Esperamos, asimismo, haber puesto de relieve el enriquecimiento que supone el uso de herramientas estadísticas en el ámbito de las investigaciones sociolingüísticas, aún más, si cabe, si son acompañadas de la computación, pues en conjunto permiten tratar de manera sistemática una cantidad ingente de datos que permitirán extraer conclusiones más precisas y menos intuitivas, conducentes al establecimiento de teorías y comprobaciones empíricas de hipótesis.

Finalmente, confiamos en haber transmitido la importancia de la interdisciplinariedad para los investigadores en sociolingüística, fomentando y potenciando los lazos con otros campos del saber y hasta creando lazos nuevos que hayan podido pasar, hasta ahora, desapercibidos u obviados, y que, sin embargo, podrían nutrir la labor investigadora de los sociolingüistas, así como la confirmación de que el matrimonio entre las ciencias y las letras no solamente es posible, sino necesario.

Referencias bibliográficas

Casas Sánchez, J. M., y Gutiérrez López, P. (2011). *Estadística ii: Inferencia estadística*. Madrid: Editorial Universitaria Ramón Areces.

Fasold, R. (1984). *La sociolingüística de la sociedad*. Madrid: Visor Libros.

Hernández Campoy, J. M., y Almeida, M. (2005). *Metodología de la investigación sociolingüística*. Málaga: Comares.

López Morales, H. (1994). *Métodos de investigación lingüística*. Salamanca: Ediciones Colegio de España.

López Morán, L., y Hernández Alonso, J. (2012). *Estadística descriptiva*. Madrid: Ediciones Académicas, S.A.

Mairal Usón, R., Peña Cervel, M. S., Cortés Rodríguez, F. J., y Ruiz de Mendoza Ibáñez, F. J. (2010). *Teoría lingüística. métodos, herramientas y paradigmas*. Madrid: Editorial Universitaria Ramón Areces.

Martín-Pliego López, F. J., y Ruiz-Maya Pérez, L. (2010). *Fundamentos de probabilidad* (2ª ed.). Madrid: Paraninfo.

Paiva Boléo, M. (1974). *Estudos de linguística portuguesa e românica*. Coimbra: Universidade de Coimbra.

Piergiorgio, C. (2007). *Metodología y técnicas de investigación social*. Madrid: McGraw-Hill.

Shannon, C. E., y Weaver, W. (1949). *The mathematical theory of communication*. Illinois: University of Illinois Press.